EFFICIENT STEREO VIDEO ENCODING FOR MOBILE APPLICATIONS USING THE 3D+F CODEC

Fabio Maninchedda[1,2]*, Marc Pollefeys*[1,2]*, Alain Fogel*[2]

[1]Computer Science Department, ETH Zürich, 8092 Zürich, Switzerland
[2]Nomad3D, Zürich, Switzerland and Nice, France

ABSTRACT

This paper presents a stereo video codec called 3D+F which has specifically been designed to meet the low complexity requirements imposed by mobile applications while trying to be competitive with the state-of-the-art coders, such as the multi view video coding (MVC) extension of the H.264/AVC standard, in terms of rate-distortion (RD). By exploiting the stereo geometry of stereoscopic image pairs it is possible to transcode a 3D movie into a 2D movie and a 2-bit labeling. The 2D movie can then be compressed using any video coder while the labeling needs to be compressed losslessly. Preliminary subjective testing shows that the resulting quality has potential to be very competitive.

***Index Terms* —** Stereo image processing, Stereoscopic representation, Mobile computing

1. INTRODUCTION

As stereoscopic 3D is becoming popular on mobile devices the need for new efficient technologies for the 3D video processing chain are required. This is mainly due to the fact that for mobile applications the constraint of limited power has to be considered. In particular the decoding should be of low complexity to allow the playback of longer movies without draining the battery. The transition from 2D to stereoscopic 3D leads to a doubling of the amount of data to be processed. Efforts towards improved coding efficiency of 3D video have been made in the MVC extension of the H.264/AVC standard [1, 2]. However this efficiency comes at the cost of a high complexity. To efficiently code multi-view videos it is necessary to additionally exploit the inter-view dependencies next to the temporal ones used in traditional video coding. For this purpose in MVC the prediction may choose among temporal and inter-view prediction and select the predictor which leads to the best coding efficiency on a block basis. Our approach is more ambitious and tries to remove the inter-view dependencies completely by understanding the scene geometry to detect shared regions in the image pairs. The necessary knowledge is acquired by performing a disparity estimation step for the stereo video. The disparity information is then used to go from the image pair to a so called cyclopean representation. The cyclopean representation is composed of a cyclopean view, which is an image that contains the shared regions along with the occluded regions, and a visibility labeling that for each pixel in the cyclopean tells in which view it is visible. Given the cyclopean view and the labeling it is possible to reconstruct the original frames. Instead of having to encode a stereoscopic video one has to encode a 2D video and a labeling.

2. OVERVIEW OF THE 3D+F PIPELINE

The encoding of a stereo movie using the proposed approach involves the following steps (see Figure 1). In an optional preprocessing step the input movie is rectified and color corrected. Rectification is required as it is assumed that corresponding pixels lie on the same scanline. Color equalization is recommended as it improves the disparity compensated prediction. Then, given the left and right input images L_1 and R_1 a disparity estimation step is performed to compute a disparity map D. From the input images and the disparity map one can compute the cyclopean view C and the corresponding 2-bit visibility labeling V. The cyclopean view is a 2D image and can therefore be encoded using any video coder while the visibility labeling needs to be encoded using a lossless entropy coder. Finally the cyclopean video and encoded labeling are multiplexed into an output stream which is denoted by S. The

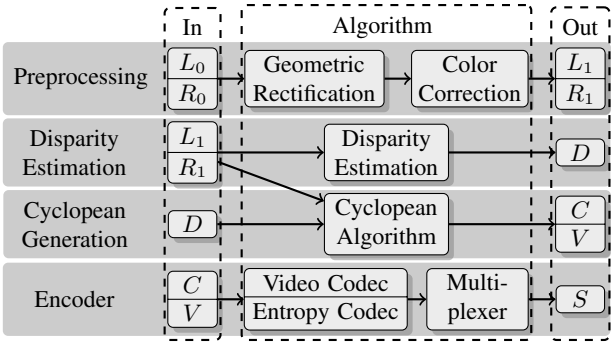

Figure 1: Encoding pipeline.

decoder works as follows (see Figure 2). The output stream S is demultiplexed and decoded to obtain the cyclopean frames C' and the corresponding labeling V. Due to the lossy video coding $C' \neq C$. Then the cyclopean frames are decoded to obtain the decoded left and right views L'_1 and R'_1. Note that the transition to a cyclopean representation and back is also a lossy process.

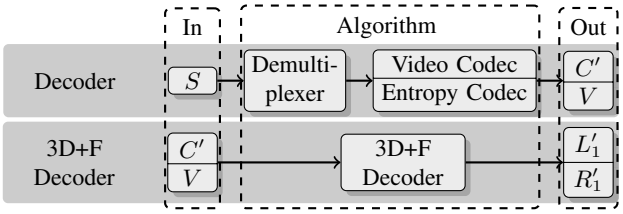

Figure 2: Decoding pipeline.

3. PREPROCESSING

The preprocessing steps are required for non properly calibrated movies and can be skipped whenever the data to be coded meets all the assumptions made in the 3D+F algorithm. Geometric rectification is required to obtain competitive results due to the assumption that corresponding pixels lie on the same scanline. Color correction should be applied whenever the mismatch of the pixel intensities is significant as it will negatively affect the PSNR of the encoded sequences. In the remainder of this section the geometric rectification and color equalization procedures are briefly explained.

3.1. Geometric Rectification

The geometric rectification is performed using the method presented in [3]. Instead of rectifying each frame independently it is assumed that the intrinsic camera parameters are constant during each scene. This model can be extended to handle scenes with time-varying camera intrinsic parameters which are currently not handled. A possible way to handle such sequences would be to rectify each frame independently while enforcing some temporal consistency to avoid inconsistencies among successive frames. Since the method has not been specifically developed for data that is used for stereoscopic viewing purposes possible improvements include the minimization of an error function which leads to the least possible amount of visual distortion. For almost rectified stereo pairs, however, it is expected that the introduced distortion is small. To obtain optimal results one should also perform a radial distortion estimation.

3.2. Color Equalization

First correspondences are computed using stereo matching (see Section 4). Then the brightness transfer function (BTF) is computed from the joint histogram using a robust maximum likelihood estimation for each color channel independently [4]. The BTF can be used to map the intensities of one image to the other. Example histograms before and after color correction using the proposed method are shown in Figure 3. This method does not account for inter channel dependencies as every color channel is treated independently. More involved approaches that model those dependencies may lead to improved results.

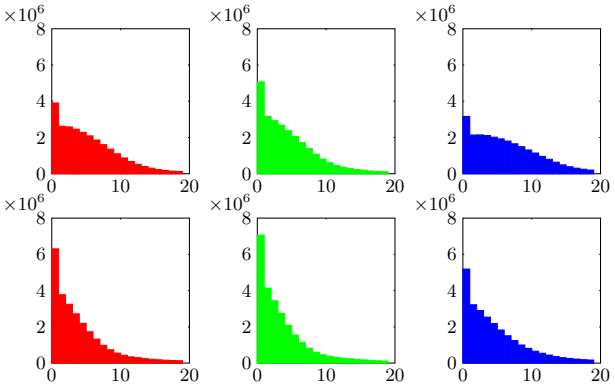

Figure 3: Error histograms before and after color correction for R, G and B color channels.

4. DISPARITY ESTIMATION

The problem of estimating disparities has been studied extensively in the past [5]. In our application scenario the goal is slightly different, in particular, the goal is to recover disparity information which leads to the best possible prediction of the right view from the left one while minimizing the size of the cyclopean. The two goals are partially in conflict because in general it is not true that the image of corresponding pixels represents the best possible match in terms of color differences. This is due to variations in gain and bias and effects such as different sampling, specular highlights and others. In general, a piecewise constant disparity map will lead to a smaller cyclopean representation, which again implies that one cannot simply choose the disparity assignment which locally minimizes the color difference of the corresponding pixels because that would lead to a quite random result. The problem of estimating such a disparity map is less ambiguous than the traditional stereo goal. Indeed, untextured regions, which are difficult to match in the traditional stereo matching problem are trivially handled in our case, since most disparity assignments will provide a result that fulfills our goals. Even though enforcing more smoothness can reduce the size of the cyclopean, it is essentially dominated by the amount of occlusions in the stereo pair. Currently the disparity estimation is performed using the method presented in [6]. In the remainder of this section an extension to improve the temporal consistency is presented. This aspect has mostly not been taken into account by currently available disparity estimation methods. The presented approach follows the lines of those methods which use the motion between adjacent frames to impose some additional constraints in the stereo matching process [7]. Assume that for two consecutive left views the motion field $\mathbf{m}^t$ that maps points at time t to corresponding points at time $t+1$ is known, in formulas $\mathbf{p}^{t+1} = \mathbf{p}^t + \mathbf{m}^t_{\mathbf{p}}$. Such a motion field can be computed using optical flow methods [8]. Then, to enforce some temporal continuity for the disparities of the corresponding pixels in time $D^t(\mathbf{p}^t)$ should not differ substantially from $D^{t+1}(\mathbf{p}^{t+1})$ whenever assigning the same disparity appears to be appropriate. In essence, if $\|\mathbf{m}^t_{\mathbf{p}}\| \approx 0$, which is the condition for which the point $\mathbf{p}^t$ is stationary, the assignment of a different disparity should be penalized. However, if there is substantial motion the disparity could change by a few values and the penalty should therefore be much lower. To achieve this goal the energy functional that is minimized in the stereo matching is extended by a consistency term of the form

$$P_T \cdot \left(1 - \exp\left\{-\frac{|D^{t+1}(\mathbf{p}^{t+1}) - D^t(\mathbf{p}^t)|}{\gamma \cdot \|\mathbf{m}^t_{\mathbf{p}}\|}\right\}\right) \qquad (1)$$

where P_T is a user-defined penalty term and γ influences the strength of the motion magnitude on the penalty. In particular, the larger γ the smaller the penalty also for low motion. The penalty can never exceed P_T, such that if the motion estimation fails the stereo optimization has still some freedom to choose a better match. Furthermore, it is known that optical flow methods tend to become inaccurate for large motion. Therefore, the adaptation of the cost function to the motion magnitude seems to be a reasonable choice since it can handle this type of inaccuracies properly. Nonetheless, the choice of reasonable values for parameters P_T and γ is crucial. Indeed, if too strong constraints are enforced the disparity map is essentially warped according to the motion field.

(a) Left (b) Right (c) Disparity (d) Cyclopean (e) Labeling

Figure 4: Illustration of the cyclopean (d) and corresponding labeling (e) for input stereo pair (a,b) and disparity (c). Label V_B is black, V_L is dark gray and V_R is light gray.

5. CYCLOPEAN GENERATION

The fundamental insight is that both views share a considerable amount of information. Indeed, most parts of the image represent the same scene points. Scene points that are not visible from one viewpoint while being visible from the other are called occluded pixels. In the following discussion it is assumed, without loss of generality, that the left view is the reference view. The cyclopean view consists of an image which contains all pixels of the left view along with its occluded pixels, plus a labeling which stores the visibility of each pixel. It is constructed row by row and the occluded pixels are inserted at the position where the occlusion occurs. The labeling can assume one of the three values $V \in \{V_B, V_L, V_R\}$, where V_B denotes that the pixel is visible in both views, V_L denotes that it is visible in the left view only and V_R denotes that it is visible in the right view only. In Figure 4 it is possible to see a stereo pair and the corresponding cyclopean representation. In the cyclopean representation some of the spatial and temporal consistency present in the original imagery is lost lost due to the misalignment of adjacent scanlines. To recover, at least partially, the spatial and temporal consistency it is possible to perform an alignment step. Please note that the conversion to a cyclopean representation and back is a lossy process. Indeed, corresponding pixels are stored only once even though they might be slightly different due to color mismatches and different sampling.

6. ENCODER AND DECODER

The cyclopean stream is a 2D movie and can therefore be encoded using any video coder. On the other hand the labeling needs to be encoded losslessly and for this purpose a CABAC [1] style encoder has been used. The prediction of the current label at time t is based on the labels in a small neighborhood that have already been encoded and the previously encoded label at the same position at time $t-1$. Decoding cyclopean frames is trivial as it requires a simple lookup of the visibility labeling.

6.1. Encoding Complexity

The most expensive step in the encoding pipeline is the disparity estimation with a complexity of $O(|\mathcal{D}|wh)$, where $|\mathcal{D}|$ denotes the disparity range and w and h the width and height of the views respectively. The motion estimation step used by video coders has a higher complexity than the proposed disparity estimation step. Due to the fact that the cyclopean movie is just slightly larger than one of the original views (10-20%), the encoding complexity is much cheaper than the one for encoding the full 3D movie.

6.2. Decoding Complexity

The decoding of the cyclopean movie is much cheaper due to the small size increase with respect to a single view. Furthermore, the complexity of the 3D+F decoder is linear in the size of the cyclopean. Therefore, the whole decoding process is very efficient.

7. EXPERIMENTAL RESULTS

All coding experiments are carried out with the H.264/AVC Reference Software JM 18.2. The encoding parameters are reported in Table 1. The test movies used for the evaluation are part of

Parameter	**Value MVC**	**Value 3D+F**
Profile	Stereo High profile	High profile
Temporal prediction	hierarchical bi-predictive	
Number views	2	1
Intra period / GOP Size	16	16
Symbol mode	CABAC	
Search range/mode	32	EPZ search
RD optimization	on	
Subpel ME	on	
I/P/B modes	on	

Table 1: Encoder configurations.

the MPEG 3DVC activity [9]. No preprocessing has been done as the sequences are properly rectified and the color mismatches are minor. The RD plot for the sequences Kendo, Dancer and Poznan Street are shown in Figure 5. The plots include the bitrates required for encoding the visibility labeling which are reported in Table 2. Currently the labeling bitrate is independent of the bitrate at which the cyclopean movie is encoded. Therefore, the overhead

Sequence	Kendo	Dancer	Poznan Street
Bitrate [kbps]	1210	1655	1696

Table 2: Bitrate of visibility labeling for various sequences.

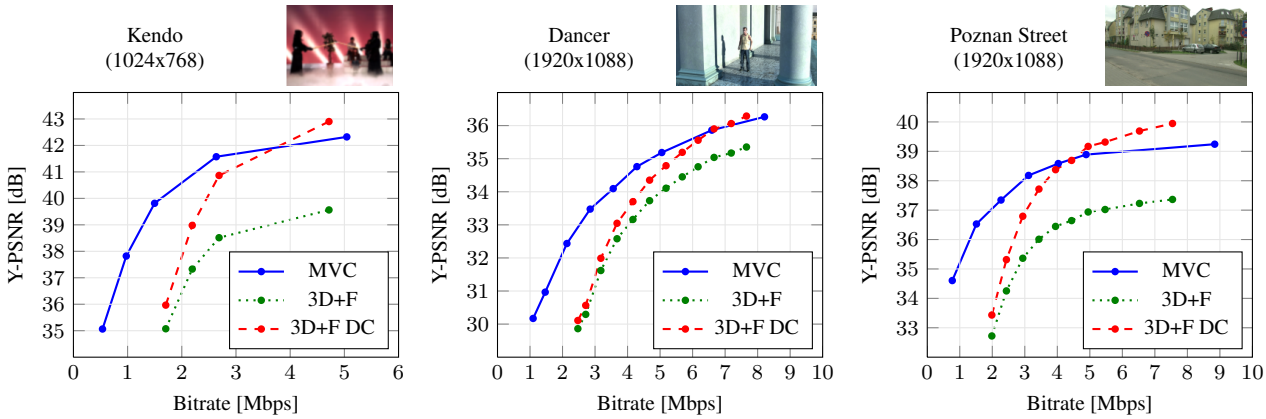

Figure 5: RD plots of MVC, 3D+F and 3D+F with PSNR evaluated on decoded cyclopean (DC) for the sequences Kendo, Dancer and Poznan Street. Both the 3D+F and 3D+F DC curves include the bitrates required for encoding the labeling which are reported in Table 2.

at low bitrates is substantial. In the future support for RD optimization between cyclopean movie and labeling will be explored. The plots in Figure 5 also show the PSNR of the 3D+F algorithm when evaluated with respect to the decoded cyclopean (DC). Recall that the cyclopean conversion introduces some distortion and therefore the 3D+F DC curve should give a hint at what visual quality one can expect assuming that the conversion to the cyclopean representation does not introduce any visual artifacts. Distortions in the cyclopean conversion are mainly introduced by color differences and sub-pixel shifts as only integer valued disparities are used. Most of this distortions are not visually significant but affect negatively the PSNR. Even though the RD performance of 3D+F is inferior to MVC according to PSNR the visual quality has potential to be very competitive. In Figure 6 the image quality at similar bitrates and similar distortions for the Poznan Street sequence are shown. It is possible to see that the quality of MVC at a similar bitrate to 3D+F is just slightly better even though there is a difference of 1.95 dB in PSNR. At similar distortion the visual quality of MVC is clearly worse than 3D+F. This can be explained by the 3D+F DC curve which at 3567 kbps has similar distortion values to MVC.

(a) (b) (c)

Figure 6: Visual comparison between MVC and 3D+F at similar bitrates or similar quality. Image (a) shows the coding results of MVC at 4925 kbps (38.89 dB), image (b) shows the coding results of 3D+F at 5263 kbps (36.94 dB) while image (c) shows the coding results of MVC at 1525 kbps (36.53 dB).

8. CONCLUSION AND FUTURE WORK

A novel low complexity stereoscopic 3D encoder has been presented. Preliminary results show that while inferior in terms of PSNR the perceived visual quality is expected to be very competitive and comparable to that of current state-of-the-art encoders such as MVC. Future work will explore ways of optimizing the RD of cyclopean movies and labeling jointly.

9. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 and ITU-T SG16/Q.6, "Draft ITU-T Recommendation H.264 and Draft ISO/IEC 14 496-10 AVC," Doc. JVT-G050, 2003.

[2] Ying Chen, Ye-Kui Wang, Kemal Ugur, Miska M. Hannuksela, Jani Lainema, and Moncef Gabbouj, "The emerging mvc standard for 3d video services," *EURASIP Journal on Applied Signal Processing*, 2008.

[3] A. Fusiello and L. Irsara, "Quasi-Euclidean Uncalibrated Epipolar Rectification," in *International Conference on Pattern Recognition*, 2008.

[4] Seon Joo Kim and M. Pollefeys, "Robust radiometric calibration and vignetting correction," *Pattern Analysis and Machine Intelligence*, 2008.

[5] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *International Journal of Computer Vision*, 2001.

[6] M. Bleyer and M. Gelautz, "Simple but Effective Tree Structures for Dynamic Programming-based Stereo Matching," in *International Conference on Computer Vision Theory and Applications*, 2008.

[7] M. Dongbo, Y. Sehoon, and A. Vetro, "Temporally consistent stereo matching using coherence function," in *3DTV-Conference*, 2010.

[8] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A Database and Evaluation Methodology for Optical Flow," in *International Conference on Computer Vision*, 2007.

[9] ISO/IEC JTC1/SC29/WG11, "Call for Proposals on 3D Video Coding Technology," Doc. N12036, 2011.